Modeling intentional agency : a neo-Gricean framework

Sarkia, Matti

2021-12

Sarkia , M 2021 , ' Modeling intentional agency : a neo-Gricean framework ' , Synthese , vol.
þÿ 1 9 9 , n o . 3 - 4 , p p . 7 0 0 3 7 0 3 0 . h t t p s : / / d o i . o r g / 1 0 . 1 0 0 7 / s 1 1 2 2 9 - 0 2 1 - 0 3 1

http://hdl.handle.net/10138/337983
https://doi.org/10.1007/s11229-021-03103-w

Synthese (2021) 199:7003–7030
https://doi.org/10.1007/s11229-021-03103-w

Modeling intentional agency: a neo-Gricean framework

Matti Sarkia[1]

Received: 22 July 2020 / Accepted: 24 February 2021 / Published online: 16 March 2021

Abstract

This paper analyzes three contrasting strategies for modeling intentional agency in contemporary analytic philosophy of mind and action, and draws parallels between them and similar strategies of scientific model-construction. Gricean modeling involves identifying primitive building blocks of intentional agency, and building up from such building blocks to prototypically agential behaviors. Analogical modeling is based on picking out an exemplary type of intentional agency, which is used as a model for other agential types. Theoretical modeling involves reasoning about intentional agency in terms of some domain-general framework of lawlike regularities, which involves no detailed reference to particular building blocks or exemplars of intentional agency (although it may involve coarse-grained or heuristic reference to some of them). Given the contrasting procedural approaches that they employ and the different types of knowledge that they embody, the three strategies are argued to provide mutually complementary perspectives on intentional agency.

Keywords Philosophical methodology · Scientific modeling · Philosophical naturalism · Philosophy of mind and action · Intentional agency

✉ Matti Sarkia
Matti.sarkia@helsinki.fi; Matti.sarkia@gmail.com

[1] Department of Political and Economic Studies, University of Helsinki, Unioninkatu 40 A 514, 00014 Helsinki, Finland

1 Introduction

This paper distinguishes three contrasting methodological strategies for modeling intentional agency in contemporary analytic philosophy of mind and action.[1] The three strategies have all been used previously, and I will provide examples of each of them from the literature. However, they have not been previously analyzed as *alternative strategies*, which possess their own distinctive constraints and affordances, by means of which models of intentional agency can be constructed. Rather, they have been understood as common manifestations of a more or less univocal philosophical practice of conceptual analysis (Jackson, 1998; King, 2016). By contrast to this traditional approach, I will argue that much like the theoretical modeling of other complex phenomena in science, such as climate change (Lenhard & Winsberg, 2010; Parker, 2006) or disease epidemics (e.g. Kermack & McKendrick, 1927; Shiller, 2017) may benefit from the use of multiple complementary modeling strategies, also philosophers can also make use of several different methodological strategies for modeling intentional agency.[2] These strategies may serve different theoretical goals to varying degrees of satisfaction, they may be more or less amenable to different types of targets, and they may complement (or sometimes, compete with) one another in our overall effort to come to terms with intentional agency. Thus philosophical research on intentional agency will be argued to resemble scientific modeling in some important respects, where different strategies of model-construction are used to capture different aspects of the causal structure of the world, and to respond to different theoretical goals and concerns (see Godfrey-Smith, 2006a; Weisberg, 2013).

The model-based methodological approach of this paper bears a similarity to meta-theoretical views that have been advanced during recent years in other domains of philosophical investigation, such as metaphysics (Godfrey-Smith, 2006b; Paul, 2012), epistemology (Williamson, 2017) and (certain parts of) the philosophy of mind (Godfrey-Smith, 2005; Maibom, 2003). For example, Williamson (2017) understands formal work on epistemic logic (especially in the Bayesian tradition) as model-construction, while Godfrey-Smith (2006b) casts David Lewis's (1986) work

[1] The notion of intentional agency is not one that I will attempt to define explicitly here, partly because of dangers of circularity that easily emerge in reductive definitions of this notion. The idea of performing purposive, or goal-directed actions is arguably central to intentional agency, and this seems to require that the notion of having intentional states (e.g. goals) is understood prior to understanding intentional agency. Many philosophers also take a kind of rational unity of perspective to be central to intentional agency, and have appealed to such unity to argue that certain kinds of aggregates of individuals (e.g. organized social groups) can be treated as agents in their own right (e.g. List & Pettit, 2011; Rovane, 1998). In this paper, I will not take a stance on whether rational unity of perspective is required for intentional agency, or whether some types of social groups ought to be recognized as agents. However, my approach is consistent with the idea that different definitions of agency may be appropriate to different types of philosophical investigation.

[2] For example, the types of fine-grained details about the nature of perceptual processes that are germane to philosophical discussions of qualia and consciousness (e.g. Bermudez 2018; Noë 2004) do not seem to be as important for the causal understanding and explanation of overt forms of behavior (e.g. Bratman, 1987; Dretske, 1988; Mele, 2009).

on possible worlds and Humean supervenience as theoretical modeling. However, by contrast to these earlier contributions, my primary concern in this paper is with contrasting *strategies* of model-construction, rather than with the meta-theoretical status of philosophical accounts (of intentional agency) as theoretical models. To be clear, my claims in this paper are restricted to the methodology of model-construction, and I do not claim that philosophical research on intentional agency is similar to scientific research in all respects. For example, philosophical investigation rarely involves empirical experimentation (although it sometimes does—see e.g. Knobe & Nichols, 2008; Michael & Szigeti, 2019), and philosophers tend to rely in their exercises of model-construction more on conceptual intuitions related to what Sellars (1963) called the *manifest image of the world* than modelers in other branches of science. This being said, I will argue that the parallels between philosophical research on intentional agency and scientific modeling are substantial enough to force us to substantially revise our conception of philosophical methodology in this domain of investigation. In particular, I will argue that a model-based perspective supports a much more *pluralistic* and *pragmatic* approach to the philosophy of mind and action than is the mainstream today.

My perspective on model-construction in science is informed by recent naturalistic approaches to the philosophy of science, which treat model-construction as one important, but not the only kind of theoretical activity that scientists engage in (Downes, 2011; Frigg & Hartman, 2012; Giere, 1988; Godfrey-Smith, 2006a; Hausman, 2012; Mäki, 2009; van Fraassen, 1980; Weisberg, 2013). According to many accounts, scientific modeling can be distinguished from other strategies of scientific theorizing by its *mediated* character (see esp. Godfrey-Smith, 2006a; Weisberg, 2007a). Instead of studying the world *directly*, by means of observation and experimentation, scientific modelers study the world *indirectly*, by first studying a model and then formulating *theoretical hypotheses* (Giere, 1988) about how (if at all) the world may be similar to the model in question.[3] The indirect nature of model-based science makes it possible for scientists to study phenomena, which could not easily be studied empirically because of causal complexity, epistemic opacity or rarity in nature. For example, social scientists have studied urban segregation with agent-based models (Schelling, 1978) and biologists have constructed models of three-sex mating to understand what constraints sexual dimorphism poses on the evolution of species (Fisher, 1930; Weisberg, 2013, pp. 131–134). There are many different kinds of models in science, from the concrete models exemplified by model organisms and scale models, to agent-based models and computer simulations, and to the abstract mathematical frameworks of ecology and rational choice theory (Downes, 2011; Frigg and Hartman, 2012). These different models exploit contrasting conceptual

[3] According to Weisberg (2007a, pp. 209–210),""Modeling… is the indirect theoretical investigation of a real world phenomenon using a model. This happens in three stages. In the first stage, a theorist constructs a model. In the second, she analyzes, refines, and further articulates the properties and dynamics of the model. Finally, in the third stage, she assesses the relationship between the model and the world if such an assessment is appropriate. If the model is sufficiently similar to the world, then the analysis of the model is also, indirectly, an analysis of the properties of the real-world phenomenon. Hence modeling involves indirect representation and analysis of real-world phenomena via the mediation of models.".

and explanatory resources, and make contrasting assumptions about the phenomena, which they are concerned with.

This paper will discuss models of intentional agency that are constructed by philosophers for the purposes of understanding and explaining (particular forms of) purposive behavior.[4] There is an important and often recognized continuity between such philosophical models of intentional agency and the "folk psychological" models that ordinary people use to make sense of each other's intentional activities (Godfrey-Smith, 2005; Lewis, 1972; Maibom, 2003; Menzies, 2010; Sellars, 1963). This continuity is in part attributable to the fact that many philosophers of mind and action work by way of systematic elaboration and elucidation of folk psychological concepts and categories, which are used as raw material for the more regimented action-theoretic frameworks that they construct. However, an increasing number of philosophers have during recent decades also drawn on various types of extended resources from empirical and theoretical science, such as cognitive psychology (e.g. Bermudez, 2018; Mele, 2009; Noë, 2004) and decision theory (e.g. Bradley, 2017; Jeffrey, 1983). Thus the continuity between philosophical research on intentional agency and scientific model-construction seems worth exploring in its own right. This paper will explore such continuities, and their relation to traditional approaches (based on conceptual analysis) to the philosophy of mind and action.

To use an apt spatial metaphor, we may describe the three strategies that I will discuss as approaching the complex phenomenon of intentional agency from different directions. The research strategy that I will describe as *Gricean modeling* involves a "bottom-up" construction of the phenomenon of intentional agency from its constituent, primitive parts. The research strategy of *analogical modeling* involves a "horizontal" shift between two or more categories of presumptive agents. And the research strategy of *theoretical modeling* involves a "top-down" application of some suitably abstract and domain-general theoretical framework to the domain of agential behavior. The three strategies also embody different types of knowledge. Gricean modeling involves an *engineer's knowledge* by opening the black box of the mind to reveal how each primitive building block of intentional agency contributes to the purposive behaviors of an agential system. Analogical modeling is typically based on (propositionally expressed) *knowledge by acquaintance* of familiar kinds of intentional agents and their possible structural similarities with other agential types. And theoretical modeling embodies a *theoretician's knowledge* of abstract theoretical generalizations that are applicable to an open-ended domain of agential phenomena. Given the different methodological strategies that they employ and the different types of knowledge that they embody, I will argue that the three strategies complement one another in our overall goal to understand intentional agency.

[4] Some representative examples of the kind of work that I have in mind are Bratman (1987, 1999, 2014), Harman (1986), Mele (2009), Searle (1983), Tuomela (2013), and Velleman (2009). Unfortunately, I will not be able to discuss their substantive views of intentional agency in detail, as my primary concern is with *procedural* differences between different strategies of model-construction, not with the types of models that have been produced by these strategies.

Taken together, the three strategies can be seen to constitute a comprehensive, "neo-Gricean" (I will explain the reasons for this label in the next section) framework for modeling intentional agency in contemporary philosophy of mind and action. However, some caveats about the scope of this paper are also in order. First, I recognize that there are further strategies for reasoning about intentional agency in philosophical contexts, including ones that do not draw on a distinctive strategy of model-construction at all—for example, ones based on phenomenological introspection (e.g. Dreyfus, 2007; Schmid, 2014; cf. Gallagher and Zahavi, 2008, pp. 153–171). Second, various amalgams of the three strategies may exist, and some philosophers may use several of the three strategies in parallel, without necessarily being aware of their important procedural differences.[5] Third, I will focus on philosophical issues that relate to the causal understanding and explanation of intentional agency, and I will largely sidestep issues about the normative justification of actions that arise in the context of moral and political philosophy (although such issues would be worth a more extensive treatment on their own). Fourth, I will also leave for another occasion discussion of model-based interdisciplinary exchanges (Grüne-Yanoff & Mäki, 2014) between philosophy and other disciplines, which are concerned with intentional action and agency, such as decision theory (Binmore, 2009; Gintis, 2009; Hausman, 2012) and social and developmental psychology (Carpenter & Svetlova, 2017; Gopnik & Meltzoff, 1997; Rakoczy, 2017; Tomasello, 2019). With these caveats in mind, I am confident that the methodological framework that I will articulate is sufficiently comprehensive to support an important kind of methodological pluralism in the philosophy of mind and action.

The first section of this paper is concerned with the bottom-up research strategy of Gricean modeling. After introducing Grice's seminal methodological program and the contrast that he drew between model-construction and conceptual analysis, I will discuss its relation to rational interpretation, conceptual engineering, and (by extension) the use of simulation methods in science. In the second section, I will turn to the "horizontal" research strategy of analogical modeling, using research on analogical inference in cognitive science to analyze the tacit principles underlying this methodological strategy, and passages from Raimo Tuomela's work on group agency to illustrate its use in the philosophy of mind and action. In the third section, I will discuss the top-down research strategy of theoretical modeling by reference to the Lotka-Volterra model of predator–prey interactions in ecology and the research program of common sense (or "analytic") functionalism in the philosophy of mind and action. At the end of the paper, I draw together my conclusions, and indicate how they support a pragmatic and pluralistic approach to the philosophy of mind and action.

[5] For example, Raimo Tuomela, whose work on group agency I will use to illustrate analogical modeling, arguably also uses other strategies of model-construction.

2 Gricean modeling

The methodological framework for modeling intentional agency formulated in this paper is called *neo-Gricean*, because Paul Grice (1974–1975) was arguably the first contemporary philosopher to suggest a model-based approach to the philosophy of mind and action. However, Grice's methodological ideas were far ahead of their time with respect to the philosophical landscape that he occupied, as there was little discussion of model-construction as a distinctive *strategy* of scientific investigation (see Godfrey-Smith, 2006a; Weisberg, 2007a, 2007b) at the time when he was writing. Perhaps for this reason, they have remained little explored to this day. The goal of this section of my paper is to excavate the largely forgotten insights of Grice's methodological program, and to use them as a foundation for constructing the extended neo-Gricean framework that I will formulate in this paper. However, my primary aim is not to carry out an exegetical study of Grice's philosophy—rather I will interpret Grice's methodological contributions liberally, drawing a number of parallels between it and certain strategies of model-construction that are used in contemporary science, and defending a type of methodological pluralism in the philosophy of mind and action that Grice did not explicitly argue for (see e.g. Grice, 1974–1975, pp. 36–37). This being said, it is useful to begin by considering how Grice (1974–1975) introduced his seminal methodological program in his Presidential Address to the 49th Annual Meeting of the American Philosophical Association:

> One procedure, which I do not in the least despise, would be to take as a body of data our linguistic intuitions concerning what we would or would not be prepared to say when using psychological terms… I am, for various reasons, not happy to confine myself to these methods… The method which I should like to apply is to construct (in imagination, of course), according to certain principles of construction, a type of creature, or rather a sequence of types of creature, to serve as a model (or models) for actual creatures.. My creatures I call pirots… The general idea is to develop sequentially the psychological theory for different brands of pirot, and to compare what one thus generates with the psychological concepts we apply to suitably related actual creatures, and when inadequacies appear, to go back to the drawing-board to extend or emend the construction (which of course is unlikely ever to be more than partial). (Grice, 1974–1975, p. 37)

The passage presents model-construction as an alternative to conceptual analysis, as it was practiced in the context of analytic philosophy of mind and action at the time when Grice was writing. What was the methodological contrast that he had in mind? The reference that Grice made to the use of "linguistic intuitions concerning what we would or would not be prepared to say when using psychological terms" suggests that he took the methods of ordinary language philosophy (which was by that time waning in influence) as the primary foil for his methodological proposal. As is well known, Grice was a vocal critic of ordinary language philosophy as a comprehensive research program, although he was acutely aware of

the significance of ordinary thought and talk for many domains of philosophical discourse (Grice, 1989; Neale, 1992). Thus Grice's contributions to the philosophy of mind and action were an extension of broader ideas that animated his philosophical thought.

The tradition of ordinary language philosophy has today been discredited in many domains of philosophical investigation (Forguson, 2001). However, some vestiges of the type of approach that it exemplified still seem to play a role in contemporary philosophy of mind and action. For example, an ordinary language philosopher might have argued that the concept of belief is such that one would not be warranted to say that one believes what one does not take to be true. Today, it is still not uncommon to hear philosophers appeal to linguistic intuitions about what it means for an action to be "intentional", for an agent to act "intentionally", or for a mental state to be an "intention" (see O'Brien, 2015). Of course, there may well be a role for such linguistic intuitions in philosophical investigation, if they are appropriately harnessed, and subjected to suitable criticism and revision (Cappelen, 2018; Machery, 2017). However, by questioning the authority of conceptual intuitions as the privileged means of reasoning about intentional agency in philosophy, Grice's methodological program was far ahead of its time, and still has lasting appeal.

The central problem in relying on conceptual intuitions in many domains of philosophical investigation has to do with the inherently conservative and indecisive nature of our conceptual intuitions about many issues of philosophical importance (see Grice 1974–1975, p. 36). In particular, many of our intuitions about central action-theoretic concepts, such as the concept of acting intentionally, are vague or applied in a context-sensitive manner depending on the circumstances in which they are used (Godfrey-Smith, 2005)—for example, depending on whether our concern is with predicting action or with attributing responsibility (Knobe, 2010), or whether we are talking about (more or less) fully socialized human individuals or about less "paradigmatic" agents, such as non-human animals, pre-linguistic infants, or group agents (Michael & Szigeti, 2019). Due to its concern with departing from vague and conservative philosophical intuitions, Grice's methodological approach can be seen to bear an affinity to recent approaches to philosophical methodology, which describe philosophical practice in terms of *conceptual engineering* of new or revised concepts, rather than in terms of (conservative) conceptual analysis of the "essential" meanings of existing concepts (Cappelen, 2018; Machery, 2017; cf. Grice, 1974–1975, p. 38). Given this contemporary perspective on philosophical practice, which can be traced back to Carnap's (1950) account of *conceptual explication*, philosophical analysis (understood as conceptual engineering) may arguably sometimes serve as one important *means* of constructing (propositional) models of intentional agency, rather than an alternative to it. However, conceptual intuitions should be granted at most a heuristic and defeasible role in the process of model-construction, whose results are ultimately justified by their pragmatic or theoretical usefulness in representing particular types of agents.

The methodological alternative to conservative conceptual analysis that Grice proposed was itself partially based on the ideal of an engineering science (Grice used the metaphors of a "genitor" and an "engineer" working together and posing mutually complementary goals and constraints on the process of model-construction).

The *modus operandi* of his program of creature construction was the recognition of specific adaptive challenges coming from the environment of simple imaginary agents (which he called "pirots" or "operants") and the step-by-step construction of appropriate psychological capacities providing adequate behavioral responses to such challenges. This resulted in a sequence or "hierarchy" of psychological types of increasing complexity, each building upon (and going beyond) the capacities of their ancestors in the sequence of pirot-types. While Grice placed human-like or "rational" creatures at the top his hierarchy, this was a function of the specific questions that he wanted to address, rather than an expression of anthropomorphic prejudice or narrow-mindedness—if Grice's goal had been to understand how different types of creatures (perhaps amoebas or crustaceans) could have emerged from a gradual succession of smaller steps, he would have constructed an alternative sequence of pirot-types, with different properties exhibited by the creatures at the top of his hierarchy. This is how Grice characterized the principles of creature construction:

> …The mode of construction is to be thought of as being relative to some very generally framed 'living-condition' concerning the relation of a pirot to its environment; the operations the capacity for which determines the type of the pirot are to be those which, given the posited condition, constitute the minimum which the pirot would require in order to optimize the chances of his remaining in a condition to perform just those operations. I have in mind such a sequence as operants which don't need to move at all to absorb sources of energy, operants which only have to make posture changes, operants which, because the sources are not constantly abundant, have to locate those sources, and (probably a good deal later in the sequence), operants who are maximally equipped to cope with an indefinite variety of physiologically tolerable environments (i.e., perhaps, rational pirots).

The goal of Grice's program was to understand *actual* creatures by creating an *imaginary* process that imitated selective aspects of the historical and evolutionary processes that had produced them. Although his procedure was in many ways abstract and idealized relative to actual processes of evolution by natural selection, it can be seen to have included the three features of variation, transmission and differential fitness, which are required for evolution to take place (Lewontin, 1970). First, his pirots were endowed with different adaptations to the changing environment in order to ensure their survival. Then these adaptations were passed on to the next generation of pirot, who inherited the core features of their psychology from their ancestors in the lineage of pirot-types. Finally, each generation of pirots was placed in an environment with other pirot types, who competed for scarce resources for sustenance and reproduction. Of course, Grice's creatures did not emerge over historical time, and Grice imposed more order on his sequence of pirot-types than could be located in the often erratic (Jacob, 1977) and sometimes saltatory (Gould, 2002) process of natural selection. These abstractions and idealizations were warranted, because

Grice's primary goal was to understand the actual creatures that had been produced by evolution by natural selection, rather than to reproduce all details of the historical and evolutionary processes that had produced them.[6]

By extension, we may identify an affinity between Grice's methodological program and the use of simulation methods in contemporary science (Macleod & Nersessian, 2017). A sequence of creatures resembling Grice's pirots might in principle be produced by a computer simulation, in which successive generations of pirots would co-evolve together with increasingly complex environments (populated by other pirot-types) at each stage during which the simulation is run (cf. Gardner, 1970; Wolfram, 2002). This "simulationist" interpretation of Grice's program of creature construction brings to mind agent-based simulations in other disciplines, such as computer science and sociology, where artificial agents are endowed with simple behavioral rules, and the macro-level outputs that these rules can generate are then studied through simulations from contrasting initial conditions, with the possibility of subsequently adjusting the behavioral rules or the initial conditions so as to test out alternative trajectories through state-space (Epstein, 1999; Gardner, 1970; Wolfram, 2002). For example, in Thomas Schelling's (1978) well-known checkerboard model of racial segregation, agents have a mild preference for staying next to agents of their own type, and they move around on a two-dimensional grid depending on whether this preference is satisfied. According to many philosophers of science (e.g. Humphreys, 2004; Winsberg, 2003), the central epistemic benefit of such agent-based modeling is the possibility of simulating outcomes that are not derivable from theory alone and the possibility of surprise—for example, we might find out that strongly segregated "neighborhoods" can emerge despite the absence of any explicitly "racist" attitudes in the population (Schelling, 1978; Ylikoski, 2014). Of course, this simulationist interpretation of Grice's methodological program goes beyond what can be extracted from his seminal writings dating back to the 1970s, as Grice was not in a position to anticipate the revolution in computational techniques that would come about in the following decades. However, given that a computational implementation of the process of creature construction would further insulate it from the types of conservative conceptual intuitions that Grice was suspicious of, I believe that it would have been congenial to his concerns.

The profound originality of Grice's methodological program can be thrown into sharp relief by contrasting it with other approaches to the philosophy of mind and action of his time, which viewed the observer-dependent and language-relative interpretation

[6] More precisely, these can be understood as instances of what Weisberg (2007b, 640) calls Galilean idealization, which is "the practice of introducing distortions into theories with the goal of simplifying theories in order to make them computationally tractable". Weisberg distinguishes Galilean idealization from *multiple models idealization*, which "is the practice of building multiple related but incompatible models, each of which makes distinct claims about the nature and causal structure giving rise to a phenomenon" (Weisberg, 2007b, p. 645), and *minimalist idealization*, which "is the practice of constructing and studying theoretical models that include only the core causal factors which give rise to a phenomenon " (Weisberg, 2007b, p. 642). While Weisberg distinguishes different types of idealization by their *representational ideals*, others have distinguished idealization from abstraction in terms of whether they distort the phenomena to be modeled by removing irrelevant features (abstraction) or by misrepresenting the features that are included in the model (idealization) (e.g. Thomson-Jones, 2005).

of behavior "as intentional" as the central principle guiding all philosophical investigations of action and agency (Anscombe, 1958; Davidson, 1980; von Wright, 1971). Giving voice to this view, Donald Davidson (1973) wrote in his hallmark essay *Radical Interpretation* that "if we cannot find a way to interpret the utterances and other behaviour of a creature as revealing a set of beliefs largely consistent and true by our standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything". By contrast to Davidson's (1980) *principle of charity*, which seemingly forced all purposive behavior to be interpreted as rational as a matter of the logic of intentional explanation alone, Grice's methodological program allowed for a specific type of engineering failure, if the psychological capacities that his creatures were endowed with failed to perform the functions that were designed for them. The measure of success was not rationality or irrationality but survival or extinction: for Grice, it was preferable to let nature (or his highly idealized simulacrum of nature) run its course, instead of attempting to dictate by conceptual intuitions alone how the products of its essentially random design ought to act and behave. Grice's methodology of creature construction was a philosophy of action for the twenty-first century, although its seeds were sown far before its truly revolutionary character could be fully appreciated.

The impact of Grice's program of creature construction on subsequent research in the philosophy of mind and action has been surprisingly limited, given the originality of his methodological insights. We may conjecture that at least three reasons have contributed to this regrettable oversight. First, as already pointed out above, there was little discussion of model-construction as a distinctive *strategy* of scientific investigation (Godfrey-Smith, 2006a; Weisberg, 2007a) at the time when Grice was writing, although models were discussed in the rather different set-theoretic sense of Suppes (1960) and Tarski (1953). Second, Grice made limited use of the methodological ideas that he presented in his own research, as it was the philosophy of language, rather than the philosophy of mind and action, which was his primary field of inquiry (Neale, 1992). Third, the central ideas of Grice's methodological program were rather densely formulated in a little explored part of his corpus, which was dotted with informal commentary on contemporaneous philosophical issues, which are challenging for anyone but the most ardent student of the history of analytical philosophy to appreciate (I have interpreted Grice's methodology liberally, and connected it to many issues in contemporary philosophical and scientific methodology, which were not central topics of debate at the time when Grice was writing). However, some philosophers, notably Michael Bratman (1987, 2014) have made extensive use of a Gricean methodology, testifying to the lasting significance of Grice's methodological ideas in contemporary philosophy of mind and action. I will return to Bratman's use of creature construction at the end of this paper.

3 Analogical modeling

The notion of analogy can be distinguished from the closely related notions of similarity (which it is an instance of) and metaphor (which is special case of analogy) by its emphasis on structural rather than material resemblance (Goldstone & Son, 2012; Holyoak, 2012; Gentner & Maravilla, 2018). Typical analogies follow the A:B::C:D

pattern (A is to B like C is to D), where the relational similarity between the two pairs may be masked by superficial similarity on the level of object correspondence (e.g. a dog chasing a man is analogous to a man chasing a dog, with the role of the chaser being played by the man in one scenario and the dog in the other scenario). To consider a famous analogy in the history of science, sound and water can be thought of as analogous phenomena, because they share a number of similar structural properties, such as "propagating across space with diminishing intensity, passing around small barriers, [and] rebounding off of large barriers" (Holyoak, 2012, 234). They are in this sense analogous phenomena, although sound and water are materially different in a number of respects—for example, water is wet, while sound is not. This is how Holyoak (2012, p. 234) characterizes analogical thinking:

> Two situations are analogous if they share a common pattern of *relationships* among their constituent elements, even though the elements themselves differ across the two situations… Typically one analog, termed the *source* or *base*, is more familiar or better understood than the second analog, termed the *target*…. This asymmetry in initial knowledge provides the basis for analogical transfer—using the source to generate inferences about the target. (Holyoak, 2012, p. 234)

The inferential principles underlying analogical reasoning can be further elucidated in terms of the *structure-mapping* theory of Dedre Gentner (1983a, b; Gentner & Maravilla, 2018). According to Gentner, three central principles guiding analogical inference are the preservation of relationships, systematicity and one-to-one correspondence. According to the *preservation of relationships*, information about relations between objects (e.g. electrons revolve around the nucleus like planets revolve around the sun) is preferentially processed over information about object attributes (e.g. the sun is yellow). According to the *systematicity principle*, more coherent and mutually constraining relations (e.g. distance, attractive force, being more massive than and revolving around a heavier body in the case of the Bohr model of the atom) are preferentially processed over relations that exhibit a lesser degree of systematicity (e.g. being hotter than). According to *one-to-one correspondence*, each (group of) element(s) in the source analog corresponds to one (group of) element(s) in the target analog (e.g. electrons are the planets and the sun is the nucleus). In addition to these three structural principles, the disposition to engage in analogical reasoning has also been shown to depend on the goals of the reasoner (Dunbar, 2001; Holyoak & Thagard, 1997; Spellman & Holyoak, 1996) and on context, including the framing of the task and prior training in laboratory experiments (Markman & Gentner, 1993).

The process of analogical reasoning can often be usefully thought of as proceeding in stages (Gentner & Maravilla, 2018; Gentner & Smith, 2012; Holyoak, 2012). Four of the most important stages in analogical reasoning are retrieval of the source and target analogs, structural mapping between the source and the target, evaluation of the posited structural correspondence, and learning from analogy. In the retrieval stage, the source analog (e.g. the solar system) and the target analog (e.g. the structure of the atom) are brought to working memory. While this may seem relatively trivial when explicitly instructed what the source and the target are, the crucial stage

in solving new problems is often finding an appropriate, previously unexplored source analogy (Dunbar, 2001; Gick & Holyoak, 1980).[7] In the mapping stage, each (group of) element(s) in the source analog is aligned with a corresponding (group of) elements in the target analog (e.g. the electrons as the planets and the nucleus as the sun). At this stage, one may also need to pay attention to *negative analogies*, or ways in which the source system and the target system are dissimilar, and *neutral analogies*, or ways in which possible similarity between the source and the target system is left open for further investigation (Hesse, 1966). In the evaluation stage, one assesses whether the analogy has been successful, or whether some parts of the posited structural correspondence fail to hold. If the analogy is perceived as unsuccessful, one may attempt to fix the analogy by re-calibrating the elements in the source analog and the target analog or by selecting a different source analog, or one may give up the enterprise of analogical modeling altogether in favor of some alternative theoretical approach. To the extent that the analogy is the analogy is perceived as successful, one can use it to draw further (defeasible) inferences about the target system. Typically, evaluating the correctness of a posited analogy involves empirical investigation.

The potential to learn from analogical inference is often the most important rationale for using analogical modeling in scientific contexts (Bailer-Jones, 2008; Gilboa et al. 2014; Haig, 2013; Niiniluoto, 1988), although analogical thinking in naturalistic settings is sometimes intuitive or automatic (Dunbar, 2001; Holyoak & Thagard, 1997). Research in cognitive science has shown that analogical thinking can be especially useful when confronted with novel problems and in the early and creative stages of scientific discovery (Gentner, 2002; Gick & Holyoak, 1980; Nersessian, 1999). This is because analogical reasoning generates alternative systems of hypotheses about the general principles governing the behavior of a system. Crucially, such hypotheses come in the form of structured knowledge concerning properties, which are known to co-occur in systems that are (by hypothesis) structurally similar to the target phenomenon. Tapping into such structured knowledge has an obvious advantage in terms of speed and efficiency over attempting to formulate hypotheses about the behavior of the target system from scratch, without a structured analogy to draw upon (Holyoak & Thagard, 1997). Thus analogical modeling can often be both more economical and more productive than other strategies of model-construction, especially when our knowledge of the target system is poor or incomplete.

The role of analogies in science has been especially prominent during times of conceptual change, when a new scientific paradigm or theory is about to replace an existing one (Gentner, 2002; Nersessian, 1999; Thagard, 1992).[8] Dramatic conceptual (and ultimately empirical) revolutions in science were facilitated by analogies

[7] For example, in Gick's and Holyoak's (1980) study, students were aided in their reasoning about how to cure a cancerous tumor when they had an analogous story about a group of soldiers surrounding a castle to draw upon (to cure the cancer without killing the patient, the tumor had to be zapped with weak rays from multiple directions instead of a single high-intensity ray).

[8] Some philosophers (e.g. Levy & Currie, 2015) have suggested that all scientific modeling involves analogical thinking in the sense of evaluating similarity relationships between a model and the world (see also Giere, 1988). My concern in this section is with analogical reasoning in the process of model-*con-*

between sound and water waves, the behavior of gases and billiard balls in motion, natural selection and artificial selection, and the mind and a computer. Of course, not all analogies ultimately turn out to be successful. For example, the wave analogy of sound, which was already discovered in Antiquity, turned out to be enormously productive by pointing to a range of appropriate structural similarities between sound and water waves. By contrast, the rival particle analogy of sound was eventually rejected as unfeasible (Holyoak, 2012, p. 235). In addition to their role in scientific discovery and conceptual change, analogies can also sometimes play an important role in theory confirmation and testing. For example, use of model organisms in biology and medicine seems to rely on the analogical inference that if a model organism (e.g. a mouse) responds in a certain manner to an experimental intervention (e.g. vaccination), also the relevant target organism (e.g. humans) will respond in a similar manner (Ankeny & Leonelli, 2011; Nelson, 2018; cf. Levy & Currie, 2016). This type of analogical inference can be made more reliable by empirical knowledge concerning physiological mechanisms underlying analogical responses in the two organisms (Bechtel & Richardson, 2010; Craver & Darden, 2013; Glennan, 2017; Machamer et al. 2000).

The process of analogical reasoning has not often been discussed as an explicit and systematic strategy of philosophical investigation. To be sure, many philosophers have used analogical reasoning in their argumentation, as when Wittgenstein (1953) drew an analogy between the rules of language use and the rules of a game, and Austin (1962) drew an analogy between speech acts and physical acts. Moreover, many famous philosophical thought experiments, such as John Searle's (1980) Chinese Room -argument and Ned Block's (1978) Nation of China -argument against causal role functionalism, seem to rely on analogical thinking. However, philosophers have rarely made explicit the inferential principles that analogical reasoning relies upon (however, see Niiniluoto, 1988). This may be because many philosophers continue to think of analogies as little more than spurious and informal aids to reasoning (Hempel, 1965, pp. 433–439; cf. Hesse, 1963), which may function at best in the "context of discovery", rather than in the "context of justification" (Reichenbach, 1938). Accordingly, most philosophers tend to shy away from making it explicit when they are using analogical reasoning in their investigation.[9] One refreshing exception to this tendency can be found in Raimo Tuomela's work on group agency:

> The idea of a group agent can be based on an intuitive analogy: Analogously to intentional action (or at least a central kind of singular intentional action) by an individual agent, intentional action by a group agent… is normally based

Footnote 8 (continued)

struction, rather than with analogical thinking in the evaluation of the model-to-world relationship. Even if all scientific modeling involves analogical thinking, not all models are constructed by way of analogies.

[9] The simulationist approach to folk psychology might be offered as a counterexample to this tendency. However, in simulation theory it is ordinary folk psychological "mentalizers" (i.e. the targets of philosophical investigation), rather than philosophers themselves, who are using analogical inference. See Goldman (2008).

> on reasons for action. Analogously to an individual having to coordinate the movements of her body parts… the members of a (we-mode) group coordinate their action… both synchronically and diachronically in order to achieve group goals. Analogously to an individual agent who is committed to her intended actions, the group members are committed as a group… to the group's actions. Let us also assume along with common sense that at least some groups… can intentionally perform actions (e.g., a business company buys another one). This intuitive analogy… tells us that if a we-mode group acts as a group, its members in general… must act in the we-mode, for a group only acts through its members; and if the group acts in the we-mode, this means that a substantial amount of we-mode acting by the members occurs. (Tuomela, 2013, p. 34)

This passage occurs in the context of Tuomela's *we-mode* account of group agents, according to which group agency is supervenient on the we-mode mental states of its members. Tuomela (2013, Ch. 2) distinguishes we-mode mental states from mental states in the individualistic I-mode by three criteria, which he calls group reasons, collective commitment, and the collectivity condition. According to group reasons, we-mode group members are assumed to give up their private reasons and motives in the group context, and to take their reasons for action from what the group believes, desires and intends (Tuomela, 2013, pp. 38–40). According to collective commitment, we-mode group members are assumed to think and act as the group requires, and to be accountable to the other group members for performing their parts of the group's activities (Tuomela, 2013, pp. 40–43). According to the collectivity condition, the group's goals and other attitudes can, on conceptually necessary grounds, be satisfied for any group member if and only if they are satisfied for all of the group's members (Tuomela, 2013, pp. 43–46).[10]

The preceding passage can be analyzed in terms of the four stages of analogical reasoning that were discussed above. In the retrieval stage, Tuomela brings up the source analog of individuals and the target analog of group agents. In the mapping stage, Tuomela maps information about the relations that (he takes to) hold between the notions of reason, coordination and commitment from the source analog of individual agents to the target analog of group agents. In the evaluation stage, Tuomela infers that both individuals and "we-mode" social groups can be treated as intentional agents, even if they are materially different in many respects—for example, individuals have proprietary bodies, while social groups do not. And in the learning stage, Tuomela goes on to draw a number of further inferences about the nature of group agency on the basis of the posited analogy. For example, Tuomela (2013, pp. 51–53) argues that we-mode group agents can act autonomously and be held morally responsible for their actions. Thus Tuomela's analogy view of group agents goes through the four stages of analogy retrieval, structural alignment, evaluation, and learning from analogy.

[10] For more discussion of Tuomela's philosophical ideas, see Corlett and Strobel (2017); Heinonen (2013); Preyer and Peter (2017).

Tuomela does not claim that individuals and group agents are analogous in all respects. For example, Tuomela (2013, p. 23) refers to negative analogies between individuals and groups agents when he points out that "individual mental states are generally *intrinsically* intentional in virtue of their biological nature, whereas group agents are intentional only in a *derived, extrinsic sense* in virtue of their members' collective construction of the group as an intentional entity that they identify with". Tuomela (ibid.) also points out that his we-mode account of group agents "leaves out important conceptual and factual features of agency, such as phenomenal features (e.g., qualitative sensations) and emotions, which belong to full-blown human agency". Tuomela (ibid.) illustrates this point by saying that that "bodiless group agents do not blush when ashamed, although their members may take part in collective guilt or pride and in similar shared emotions in the we-mode". Given these and other relevant disanalogies, Tuomela (ibid.) ultimately contends that "groups can never be full-blown agents (or persons) in the flesh-and-blood sense, but at best entities that share some similar functional features with intentional human agents".

There are also numerous other examples of analogical reasoning in the philosophy of mind and action. One salient example is the analogy between a mind and a computer, which has been influential both in philosophy (e.g. Fodor, 1987; cf. Searle, 1980) as well as in the behavioral and cognitive sciences. For example, in the control-systems tradition in cybernetics (e.g. Ashby, 1957; Miller et al. 1960), the mind was viewed as an input–output machine, with distinctive (perceptual) inputs, computational (inferential) rules operating on those inputs, and (behavioral) outputs. The circular organization of goal, perception, and action was taken to govern the behavior of an intentional system in a manner that is analogous to the way in which changes in ambient temperature govern the behavior of simple mechanical systems, such as a thermostats (e.g. switching the radiator on or off depending on whether the ambient temperature is in the desired interval). The analogy between the mind and a computer had a profound impact on the development of cognitive science during the twentieth century, where the mind was framed not as a particular type of substance in the world (as in the Cartesian tradition) but in terms of the distinctive types of operations that it performs. While recent philosophers of cognitive (neuro-)science have challenged the idea that the mind can be studied independently of its material basis (e.g. Boone & Piccinini, 2016; Turner, 2018), the analogy between the mind and a computer provided the basis for an enormously fruitful research program in cognitive science during much of the twentieth century. This example accordingly illustrates how new and innovative analogies can initiate "conceptual revolutions" (Thagard, 1992), which serve an agenda-setting role for an entire scientific field or discipline.

4 Theoretical modeling

Theoretical models in science are often contrasted with concrete models and scale models, such as an airplane wing in a wind tunnel, a scaled down map of a certain geographical region, or a model organism (Downes, 2011; Frigg & Hartman, 2012). However, the primary feature that sets apart theoretical *modeling* in the sense that

I am concerned with in this article is its procedural character (cf. Godfrey-Smith, 2006a; Weisberg, 2013). Theoretical modeling involves reasoning about the laws and regularities that are associated with a particular domain of phenomena without detailed reference to either particular entities that populate that domain (as in analogical modeling) or particular mechanisms that maintain those laws and regularities (as in Gricean modeling). Theoretical modeling typically also involves a kind of top-down and holistic approach in the sense that the relevant laws and regularities are characterized in interdependent terms. For example, the notions of force, mass and acceleration are characterized in an interdependent manner by Newton's laws of motion (Giere, 1988). Here, I will first discuss the research strategy of theoretical modeling in general terms, before illustrating its use in the philosophy of mind and action by the research program of common sense (or "analytic") functionalism.

Theoretical modeling can in part be distinguished from the two other modeling strategies that I have discussed by its abstract and domain-general nature. However, what counts as sufficiently domain-general is to some extent relative to the type of phenomena that one attempts to explain. For example, physical models of simple harmonic motion have been applied to a wide range of different physical systems exhibiting oscillatory behavior, such as swinging pendulums, bouncing springs, and vibrating strings (Giere, 1988). Ecological models of predator–prey interactions have been applied to many different populations of predator and prey occupying the same ecosystems, such as sharks and squid in the Mediterranean and lynx and hare in Canada (Weisberg, 2013). And economic models of rational choice have been applied to a wide variety of systems that exhibit choice-behavior, such as individuals, time-slices of individuals, households, and business corporations (Hausman, 2012; Ross, 2010; Varian, 2010). For the philosophy of mind and action, the relevant domain is the domain of intentional agents, who can be described as having representational states, such as beliefs and desires, and as acting on their basis (Dennett, 1987).

The most important procedural difference that sets apart theoretical modeling from Gricean modeling and analogical modeling is its reliance on derivational techniques, although theoretical models need not be formulated in terms of mathematical equations (in addition, they may also involve e.g. predicate logic or other methods of linguistic derivation). Theoretical modeling involves deriving a set of quantities and/or propositions representing a phenomenon from a body of generalizations (or laws) governing the domain of phenomena that it belongs to. For example, general equilibrium theory in economics involves deriving the equilibrium price for a good from general regularities governing supply and demand, such as profit maximization, diminishing marginal utility, and the law of one price (Hausman, 1992). However, price determination in markets can also be modeled by the tools of agent-based models, for example, by simulations relating consumer behavior to the perceived rate of inflation, consumers' state of optimism about the future, or simple decision-making heuristics (e.g. pick the most familiar product) (Foley & Farmer, 2009). Adopting this type of bottom-up approach can lead to different predictions than the analytically derived outcomes predicted by general equilibrium theory, although their predictions must of course be compared to empirical data in order to be verified.

The domain-general character of theoretical modeling can in many circumstances be regarded as an epistemic virtue, given that it allows us to unify a broad range of different phenomena under a common set of laws and regularities. However, one must typically sacrifice some attention to detail in order to construct a theoretical model that is highly general and broadly applicable (Matthewson & Weisberg, 2009). Thus there is no generally applicable reason to regard either a more top-down or more bottom-up approach to model-construction as more appropriate in all circumstances. Rather, different approaches to model-construction involve contrasting trade-offs between precision, domain-generality, and empirical accuracy, which may be more or less appropriate depending on the interests of the modeler and the target phenomena that they are used to study (Levins, 1966; Weisberg, 2013). To illustrate, consider a simple version of the Lotka-Volterra model of predator–prey interactions discussed in Weisberg (2007a, p. 210):

$$dV/dt = rV - (aV)P$$

$$dP/dt = b(aV)P - mP$$

The equations represent the dynamic relationship between a predator and prey population as a function of time (t) and the size of the predator (V) and prey (P) population, the intrinsic birth rates of the two populations (r for predator and b for prey) and the intrinsic death rates of the populations (m for predator and a for prey). This simple model is in many respects abstract and idealized (Matthewson & Weisberg, 2009). For example, it considers only two populations and does not take into account population structure or alternative sources of food for the predators. Despite these abstractions and idealizations, it allows us to get a handle on some of the crucial parameters governing periodic oscillations in particular predator and prey populations in particular environments. For example, the model was used to explain the unexpected depletion of certain prey populations in the Adriatic Sea after World War II, when many fishermen expected population sizes to have increased as a result of the wartime decrease in commercial fishing (Weisberg, 2007a).

Theoretical modeling in the philosophy of mind and action can be exemplified by the research program of common sense (or "analytic") functionalism. While common sense functionalism was traditionally described as making explicit a tacit "theory of the mind" (Gopnik & Meltzoff, 1997; Lewis, 1972; Sellars, 1963), common sense functionalism can also be described as involving theoretical modeling and model-construction, because it involves *indirect* representation of agents through general law-like regularities connecting mental states to behavior (Godfrey-Smith, 2006). These regularities involve many abstractions and idealizations relative to the behavior of actual agents—for example, actual agents may fail to draw the logical consequences of their beliefs, although the requirement of logical coherence is included (for reasons of tractability) in many models of intentional agency that are based on common sense functionalism (cf. Harman, 1986). The seeds of the research program of common sense functionalism were initially sown by the seminal investigations of Wilfrid Sellars (1963), and they

were later formalized using the Ramsey sentence method in predicate logic by David Lewis (1972). In a classic textbook, David Braddon-Mitchell and Frank Jackson distil it in the following terms:

> "What then is a given mental state M, on the common-sense functionalist story? It is the state that plays the M role in the network of interconnections delivered by common knowledge about the mind. The network in effect identifies a role for each mental state, and thus, according to it, a mental state is simply the state that occupies the role definitive of it. The situation is similar to that which applies when we elucidate our concept of being a bank teller. There is a network of interconnections between inputs involving customers entering the bank, outputs involving loans approved and cash handed over, and internal connections between tellers, accountants, managers, and the like. What then is a teller? Anyone who occupies the relevant role in the network." (Braddon-Mitchell & Jackson, 2007, pp. 53-54)

Common sense functionalism exemplifies the type of top-down and holistic approach that is characteristic of theoretical modeling, as it identifies the functional roles that are played by *each* type of mental state by the *entire* network of generalizations in which it figures in "common knowledge of the mind" (Braddon-Mitchell & Jackson, 2007, pp. 52–53; see also Lewis, 1972). Common sense functionalists typically take our common knowledge of the mind to be paradigmatically expressed in natural language, and to be manifested in linguistic platitudes such as "bodily damage causes pain", "desire for beer causes behaviour that leads to beer consumption", and "belief that if p then q typically causes belief that q on learning p" (see Braddon-Mitchell & Jackson, 2007, p. 52). Braddon-Mitchell and Jackson (ibid.) classify these types of platitudes into three broad categories (originally identified by Sellars, 1963; see also Searle, 1983): *world-to-mind* regularities, which describe the causal influence that the world has on mental states, *mind-to-world* regularities, which describe how mental states cause bodily behaviors, and *mind-to-mind* regularities, which express internal interconnections between different types of mental states.

Common sense functionalism is in part motivated by the search for generality and wide scope that are characteristic of the research strategy of theoretical modeling. Thus Braddon-Mitchell and Jackson (2007, pp. 49–52) regard as one of the main benefits of the (common sense) functionalist approach that it allows for the multiple realizability of mental states by different physical substrates and thereby avoids "chauvinism" about intentional agency. For example, the operations that are characteristic of the mental state of 'believing that someone is standing behind the door' might be carried out by one set of physiological states in dogs, by another set of states in humans, and by a third set of states in intelligent robots. Accordingly, many philosophers, who draw on a common sense functionalist approach to intentional agency (e.g. List & Pettit, 2011; Pettit, 1996), have made a point of arguing that even physically quite different types of systems, including many animals, simple robots and even some social groups, can be described as intentional agents with beliefs and desires from "the intentional stance" (Dennett, 1987). This is because common sense functionalists argue that as long as a system exhibits behavioral regularities

that are characteristic of intentional agency, then it counts as an agent, regardless of its physical make-up or the nature of the mechanisms underlying its operations.[11]

The common sense functionalist approach to intentional agency has been challenged by some philosophers, who have argued that our common knowledge of the mind is not structured around the types of law-like generalizations about behavior identified by common sense functionalists. For example, simulation theorists have provided evidence that much of our understanding of other people's mental life is based on our capacity to use our own decision-making system as a *physical* model of the decision-making systems of other agents (Goldman, 2008). Maibom (2003) argues that our ordinary framework of intentional agency is structured around abstract theoretical models, rather than quasi-universal generalizations with implicit *ceteris paribus* -clauses. Hutto (2008) emphasizes the narrative aspects of folk psychological action understanding against the kinds of reductive belief-desire explanations favored by many common sense functionalists. And David Velleman (2009) capitalizes on the second-order desire for self-understanding, which builds upon the causal regularities identified by basic forms of common sense functionalism.

These criticisms indicate that the scope of common sense functionalism is likely to be more limited than many philosophers educated by Braddon-Mitchell's and Jackson's (2007) classic textbook may have assumed. However, the mistake is simply to take common sense functionalism as the *privileged* methodological framework for modeling intentional agency in the philosophy of mind and action, as there are surely more resources for action understanding that both ordinary people and philosophers may use. For example, Godfrey-Smith (2005) describes folk psychology as a *family of models*, which individuals can use flexibly in the pursuit of a wide variety of different theoretical and practical goals.

> ...It is possible that a basic part of our psychological apparatus is a facility for what we might call *model-based understanding*. This skill involves the imaginary construction of simplified structures for the purpose of understanding more complex systems. If there is such an ability, the products of some kinds of folk theorizing might have important features in common with models in the scientific sense, and *both* of these may contrast with some more traditional philosophical notions of theory. Further, once we recognize the possibility of model-based understanding—in this sense—as a distinctive psychological capacity that operates both inside and outside of scientific contexts, we can note that folk psychology could have model-like features that are not very science-like.

The heterogeneity and context-specificity of folk psychological models has important repercussions for how internally unified we should consider the types

[11] This view is expressed in an exceptionally vivid manner by Pettit (1996, p. 10): "There are regularities characteristic of beliefs and desires, regularities that dictate both the effect of certain sort of evidence on what beliefs and desires are maintained, and the effect of certain sorts of belief–desire profiles on what responses are evinced… A system will count as an intentional agent to the extent that its interactions with its environment, or at least some of its interactions, are governed by such regularities.".

of knowledge delivered by common sense functionalism to be. For example, Godfrey-Smith (2005) argues that many deeply ingrained debates about realism and instrumentalism in the philosophy of mind and action can be framed in terms of contrasting *ontological construals* of folk psychology. While both Churchland's (1981) eliminativism and Fodor's (1987) "industry-strength" realism are premised on the idea that the ontological commitments of folk psychology are to be taken quite literally—despite drawing opposing conclusions about the usefulness of folk psychology from the presumption of realism— Dennett (1987) emphasizes the pragmatic uses of folk psychology at the expense of its ontological commitments. According to Godfrey-Smith (2005), the ontological and conceptual commitments of folk psychological notions can be adjusted flexibly depending on the circumstances in which they are used and the agents whose behaviors they are used to model. For example, the condition of logical closure (i.e. that agents be capable of inferring all the logical consequences of their beliefs) may seem appropriate for some cases of explicit belief-reasoning in grown-up human agents, e.g. when solving a puzzle, but this condition would hardly seem inappropriate in the case of one's pet dog wagging its tail in the belief that its owner is behind the door. Godfrey-Smith (2005) argues that these contrasting uses of the term 'belief' do not simply involve different notions of belief (e.g. 'beliefs*' and 'beliefs**') since the criteria for attributing such mental states can be highly sensitive to context, and consequently, one would end up with an indefinite variety of different notions of 'belief'. The domain-general and deductive approach of common sense functionalism seeks to abstract away from such context-specific and pragmatic features in the pursuit of a comprehensive and internally unified framework (see e.g. Lewis, 1972, p. 256). Thus it cannot provide the whole story about how we come to understand intentional agency, even if it successfully captures certain types of linguistically scaffolded reasoning about intentional agency (cf. Apperly & Butterfill, 2009).

There is a close connection between the types of folk psychological models that ordinary people use in their daily lives and the more regimented models of intentional agency that philosophers of mind and action have constructed for understanding and explaining intentional agency, as already noted in the introduction to this paper. In one sense, the extended neo-Gricean framework for modeling intentional agency that I have formulated can be viewed as an extension of Godfrey-Smith's important insights about the pluralistic and pragmatic nature of folk psychological knowledge. While Godfrey-Smith (2005) argued that the type of folk psychological apparatus that *ordinary people* use for predicting, understanding and explaining one another's intentional activities is made up of a family of models, rather than a single, internally unified theoretical framework, I have argued in this paper that *philosophers* can use many different methodological strategies for modeling intentional agency, just like the ordinary agents, whose behaviors they seek to model. In the next section, I will indicate how the three strategies that I have discussed complement one another in our overall goal to understand and explain (different forms of) intentional agency.

5 Understanding agency by modeling

This paper has analyzed three strategies for modeling intentional agency in the philosophy of mind and action. The tacit principles underlying these strategies have been exposed, they have been illustrated with examples from the philosophical literature, and they have been connected to similar modeling strategies in science. While Paul Grice (1974–1975) can be credited with first articulating the idea of model-construction as a serious alternative to the once dominant enterprises of (conservative) conceptual analysis and rational interpretation, both Grice and his followers considered only one, distinctively bottom-up form of model-construction. By contrast, I have gone beyond Grice's seminal contributions by arguing that there are in fact several different strategies of model-construction, which philosophers of mind and action can use in their research. In this concluding section of my paper, I will indicate how the three strategies can be used in parallel to support a more pluralistic and pragmatic approach to the philosophy of mind and action than has previously been available.

To begin, I will summarize the central features of the three modeling strategies. The research strategy of Gricean modeling is based on identifying primitive building blocks of intentional agency, and building up from such basic building blocks to more complex agential behaviors (Grice, 1974–1975). Its bottom-up approach, which can be seen to bear an affinity to the program of conceptual engineering in contemporary philosophical methodology and the use of simulation methods in science, aims at an *engineer's knowledge* by opening the "black box" of the mind to investigate how each primitive building block of intentional agency contributes to the goal-directed behaviors of an agential system. The research strategy of analogical modeling is based on picking out some *exemplary type* of intentional agency, which is used as a model for other, structurally similar agential types. Its "horizontal" approach, which reminds us of the use of model organisms in biology and the role of analogies during times of conceptual change, embodies (propositionally expressed) *knowledge by acquaintance* of familiar agential types and their (possible) structural similarities with other types of agents. The research strategy of theoretical modeling is based on reasoning about intentional agency in terms of a domain-general network of law-like regularities, which involves no detailed reference to either distinctive building blocks or exemplars of intentional agency (although it may involve coarse-grained or heuristic reference to some of them). Its top-down approach involves abstract *theoretician's knowledge* of general principles applying to an open-ended domain of agential phenomena, resembling the use of abstract mathematical and propositional frameworks in disciplines, such as ecology and economics (Hausman, 2012; Weisberg, 2007a, 2007b).

The research strategies of Gricean modeling, analogical modeling and theoretical modeling can arguably play mutually complementary roles in the philosophy of mind and action. Consider the following timely parallel to science: a group of scientists studying the propagation of an incipient virus epidemic across a population might make use of abstract mathematical models in epidemiology, *and*

carry out experimental interventions on various types of model organisms, *and* use computational techniques such as agent-based modeling. Similarly, my suggestion is that philosophers can use the strategies of Gricean modeling, analogical modeling, and theoretical modeling in parallel to focus on particular types or aspects of intentional agency, and to test the plausibility of the conclusions that they have drawn from a contrasting methodological perspective. This does not mean that the results of the three strategies must converge with one another, given that the procedural principles that they are based on may point in different directions, and it may sometimes be useful to have many different models of the same phenomenon (Levins, 1966; Weisberg, 2013, pp. 103–105; Wimsatt, 2007, Ch. 6).[12] However, when the three strategies do converge on similar conclusions, they can yield more compelling insights than any of the strategies taken in isolation.

To illustrate, consider the idea that the mind can be modeled in terms of (perceptual) inputs and (behavioral) outputs, as well as operations performed on internal mental representations. This idea has been supported by analogical reasoning comparing the mind to a computer (see Sect. 3), as well as a by bottom-up Gricean ideas about how a simple organism interacts with its environment, and top-down common sense functionalist ideas about the types of "world-to-mind", "mind-to-mind", and "mind-to-world" (Braddon-Mitchell & Jackson, 2007, p. 52; Sellars, 1963) regularities that have been taken to be implicit in our ordinary framework of agency. Thus it justifiably has a central place in most philosophical models of intentional agency, as it can be rationalized on the basis of numerous different strategies of model construction. On the other hand, some assumptions that are involved in more cognitively demanding functionalist models of intentional agency, such as the requirement of logical closure, may seem too demanding for bottom-up Gricean modelers, who prefer to construct models of complex agents from simple building blocks and to introduce more demanding features only when necessary. Thus Gricean modelers would prefer to relegate this feature to more advanced creatures in the Gricean hierarchy of agent-types, while they may still agree with functionalists that a certain type of responsiveness to evidence is a central feature of all mental states that can be described as 'beliefs'.

The three strategies may also sometimes lead to different conclusions or play competing roles in our efforts to understand and explain intentional agency. For example, philosophers have offered different arguments for the reducibility or irreducibility of group agency to individual agency depending on whether they have adopted a strategy of Gricean modeling, analogical modeling, or theoretical modeling. Most philosophers, whose work can be described as involving forms of Gricean modeling, have been inclined to endorse reductionist accounts of group agents (e.g. Bratman, 2014, pp. 121–131). By contrast, philosophers making use of analogical

[12] For example, Weisberg (2007b, pp. 646–647) discusses the three different models of global circulation patterns used by the US National Weather Service to forecast the weather, and rationalizes the use of multiple models by saying that"theorists have different goals for their representations, such as accuracy, precision, generality, and simplicity.. [which] can trade off with one another in certain circumstances, meaning that no single model can have all of these properties to the highest magnitude".

(e.g. Tuomela, 2013, 34–36) or theoretical strategies of model-construction (e.g. List & Pettit, 2011, pp. 31–41) have often been tempted to treat group agents as irreducible to individual agents in at least some important respects. This being said, one should not draw too far-reaching conclusions about the reductionist bias of bottom-up forms of model-construction or the anti-reductionist agenda of top-down forms of model-construction, since such deep philosophical disagreements are underdetermined by the procedural differences that I have discussed in this article. For example, consider the manner in which Michael Bratman (2014) describes the goals of Gricean creature construction:

> The aim of creature construction is to understand more complex forms of agency by building step-wise from simpler forms of agency. We build more complex structures upon a foundation of simpler structures in ways that respond to identifiable problems and issues that arise in the context of those simpler structures. (Bratman, 2014, p. 25)

This description of the goals of creature construction may appear to mask a reductionist agenda. However, on a closer reading of Bratman's research in the philosophy of action, it becomes evident that Bratman has used Gricean methodology in order to argue for *both* reductionist and anti-reductionist ideas. For example, in his early work on the belief-desire-intention (or BDI-)-framework in the philosophy of mind and action, Bratman (1987) defended the *irreducibility* of intentions as particular types of mental states against Humean accounts, which sought to analyze all forms of intentional agency in terms of the primitive building blocks of belief and desire (Davidson, 1980; Smith, 1994). By contrast, in his more recent work on shared agency, Bratman (1999, 2014) has argued against the necessity of appealing to irreducible *we-intentions* (Searle, 2010; Tuomela, 2013) or *joint commitments* (Gilbert, 2010) in the analysis of small-scale forms of shared agency in the absence of asymmetric authority relations (which Bratman calls forms of *modest sociality*). Thus bottom-up modeling is not intrinsically and unavoidably reductionist, nor is top-down modeling intrinsically and unavoidably anti-reductionist, although they may be more or less conducive to reductionist or anti-reductionist ideas depending on the types of agential phenomena that they are used to model and the types of theoretical contexts in which they are used (concerned e.g. with the prediction, explanation, or normative evaluation of agential phenomena).

The general lesson that we can draw from Bratman's use of the methodology of creature construction is that one should not confuse the procedural strategies by means of which philosophical models of intentional agency are constructed with the substantive features of the models that are constructed by means of such strategies. My main concern in this paper has been with the methodology of model-construction, not with determining which philosophical model of intentional agency is the correct or most appropriate one (in any particular set of epistemic circumstances or relative to any particular category of agents). While we may conjecture that the possibility of multiple modeling strategies pointing in a similar direction may lead to the development of agential models that are in some sense more robust (Levins, 1966; Weisberg, 2006; Wimsatt, 1981; Woodward, 2006) than models that have been supported by an isolated strategy, the notion of robustness has itself been defined in

multiple different ways in the philosophy of science, and challenged by some philosophers as a non-empirical form of confirmation (e.g. Odenbaugh & Alexandrova, 2011; Orzack & Sober, 1993).[13] Moreover, questions of robustness have to do primarily with the evaluation of the model-to-world relationship, i.e. with evidence and realism, rather than with the methodology of model-construction. Accordingly, I will not discuss the topic of robustness in more detail in the present paper, although it is an interesting and important topic in its own right. For the same reason, I will also leave it open in this paper whether agency is something objective in the world or whether it is in the "eye of the beholder", as some philosophers in the Wittgensteinian tradition have argued (e.g. Bennett & Hacker, 2003), and whether this raises issues that are similar to or different from issues of realism that relate to the existence of middle-sized physical objects, such as tables and chairs (as contrasted with the types of objects that are postulated by fundamental physics, such as quarks or electrons (Ladyman & Ross, 2007; Niiniluoto, 1999)).

There remain many opportunities for exploring further the connections that I have drawn between philosophical research on intentional agency and scientific model-construction. However, I hope to have identified sufficient parallels between these fields of research to support an important kind of methodological pluralism in the philosophy of mind and action.

Acknowledgements This paper benefited from generous comments from Dr. Tuukka Kaidesoja and two anonymous referees for *Synthese*.

Funding Research funding was provided by a grant from the Emil Aaltonen Foundation. Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

References

Ankeny, R., & Leonelli, S. (2011). What's so special about model organisms? *Studies in the History and Philosophy of Science, 41,* 313–323.

[13] Eronen (2014) distinguishes between derivational robustness and robustness as multiple accessibility. Derivational robustness has to do with several different models with contrasting simplifying assumptions supporting the same conclusions, while robustness as multiple accessibility has to do with the same phenomenon being accessible by multiple different modalities (e.g. by the sensory modalities of vision, touch, and hearing). Woodward (2006) distinguishes between robustness as insensitivity of the results of inference to alternative specifications, robustness of derivations, robustness of measurement results, and robustness as a mark of causal as opposed to (merely) correlational relationships.

Anscombe, E. (1958). *Intention*. Blackwell.
Apperly, I., & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970.
Ashby, W. (1956). *An introduction to cybernetics*. Chapman and Hall.
Austin, J. (1962). *How to do things with words* (2nd ed.). Harvard University Press.
Bailer-Jones, D. (2008). Models, metaphors, and analogies. In P. Machamer & M. Silberstein (Eds.), *The Blackwell guide to the philosophy of science* (pp. 108–127). Blackwell.
Bechtel, W., & Richardson, R. (2010). *Discovering complexity: decomposition and localization as strategies in scientific research*. MIT Press.
Bennett, M., & Hacker, P. (2003). *Philosophical foundations of neuroscience*. Blackwell.
Bermudez, J. (2018). *The bodily self*. MIT Press.
Binmore, K. (2009). *Rational decisions*. Princeton University Press.
Black, M. (1962). *Models and metaphors. Studies in language and philosophy*. Cornell University Press.
Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science, 9,* 261–325.
Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese, 193,* 1509–1534.
Braddon-Mitchell, D., & Jackson, F. (2007). *Philosophy of mind and cognition. An introduction* (2nd ed.). Blackwell.
Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.
Bratman, M. (1987). *Intention, plans, and practical reason*. MIT Press.
Bratman, M. (1999). *Faces of intention*. Cambridge University Press.
Bratman, M. (2014). *Shared agency: a planning theory of acting together*. Oxford University Press.
Burgess, A., Cappelen, H., & Plunkett, D. (2020). *Conceptual engineering and conceptual ethics*. Oxford University Press.
Cappelen, H. (2018). *Fixing language: an essay on conceptual engineering*. Oxford University Press.
Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press.
Carpenter, M., & Svetlova, M. (2017). Social development. In B. Hopkins, E. Geangu, & S. Linkenauger (Eds.), *Cambridge Encyclopedia of Child Development* (pp. 415–423). Cambridge University Press.
Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78,* 67–90.
Corlett, J., & Strobel, J. (2017). Raimo Tuomela's social ontology. *Social Epistemology, 31*(6), 557–571.
Craver, C., & Darden, L. (2013). *In search of mechanisms: discoveries across the life sciences*. University of Chicago Press.
Davidson, D. (1973). Radical interpretation. *Dialectica, 27,* 314–328.
Davidson, D. (1980/2001). *Essays on actions and events*. 2nd Edition. Clarendon Press.
Dennett, D. (1987). *The intentional stance*. MIT Press.
Downes, S. (2011). Scientific models. *Philosophy. Compass, 6*(11), 757–764.
Dretske, F. (1988). *Explaining behavior: reasons in a world of causes*. MIT Press.
Dreyfus, H. (2007). The return of the myth of the mental. *Inquiry, 50*(4), 352–365.
Dunbar, K. (2001). The analogical paradox: why analogy is so easy in naturalistic settings, yet so difficult in the laboratory. In D. Gentner, K. Holyoak, & B. Kokinov. (Eds.), *The analogical mind: perspectives from cognitive science*. MIT Press.
Epstein, J. (1999). Agent-based computational models and generative social science. *Complexity, 4*(5), 41–60.
Eronen, M. (2015). Robustness and reality. *Synthese, 192,* 3961–3977.
Farmer, J., & Foley, D. (2009). The economy needs agent-based modelling. *Nature, 460,* 685–686.
Fisher, R. A. (1930). *The genetical theory of natural selection*. The Clarendon Press.
Fodor, J. (1987). *Psychosemantics*. MIT Press.
Forguson, L. (2001). Oxford and the epidemic of ordinary language philosophy. *The Monist, 84*(3), 325–345.
Frigg, R., & Hartmann, S. (2012). Models in Science. *The Stanford Encyclopedia of Philosophy*. Fall 2012 Edition. Edward N. Zalta (Ed.). URL = http://plato.stanford.edu/archives/fall2012/entries/models-science/
Gallagher, S., & Zahavi, D. (2014). *The phenomenological mind*. Routledge.
Gardner, M. (1970). The fantastic combinations of John Conway's new solitaire game 'Life'. *Scientific American, 223,* 120–123.

Gavetti, G., & Rivkin, J. (2005). How strategists really think: tapping the power of analogy. *Harvard Business Review* April (2005).
Gentner, D. (1983). Structure-mapping. A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.
Gentner, D. (2002). Analogy in scientific discovery: the case of Johannes Kepler. In L. Magnani & N. Nersessian (Eds.), *model-based reasoning: science, technology, values* (pp. 21–39). Kluwer.
Gentner, D., & Maravilla, F. (2018). Analogical reasoning. In L. J. Ball & V. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 186–203). Psychology Press.
Gentner, D., & Smith, L. (2012). Analogical reasoning. In V. S. Ramachandran (Ed.), *Encyclopedia of Human Behavior* (2nd ed., pp. 130–136). Academic Press.
Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.
Giere, R. (1988). *Explaining science: a cognitive approach*. University of Chicago Press.
Giere, R. (2004). How models are used to represent reality. *Philosophy of science, 71*(5), 742–752.
Gilbert, M. (2013). *Joint commitment: how we make the social world*. Oxford University Press.
Gilboa, I., Postelwaite, A., Samuelson, L., & Schmeidler, D. (2014). Economic models as analogies. *The Economic Journal, 124,* 513–533.
Gintis, H. (2009). *The bounds of reason*. Princeton.
Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.
Godfrey, S., & P. . (2006). The strategy of model-based science. *Biology and Philosophy, 21*(5), 725–740.
Godfrey-Smith, P. (2005). Folk psychology as a model. *Philosopher's Imprint, 5*(6), 1–16.
Godfrey-Smith, P. (2006). Theories and models in metaphysics. *Harvard Review of Philosophy, 14,* 4–19.
Goldman, A. (2008). *Simulating minds. The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.
Goldstone, R., & Son, J. (2012). Similarity. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning*. Oxford University Press.
Gopnik, A. (1996). The scientist as child. *Philosophy of Science, 63*(4), 485–514.
Gopnik, A., & Meltzoff, A. (1997). *Words, thoughts, theories*. MA, MIT Press.
Gould, S. (2002). *The structure of evolutionary theory*. Harvard University Press.
Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.
Grice, H. P. (1974-1975). Method in philosophical psychology: from the banal to the bizarre. *Proceedings and Addresses of the American Philosophical Association*, 48, 23-53
Grüne-Yanoff, T., & Mäki, U. (2014). Introduction: interdisciplinary model-exchanges. *Studies in History and Philosophy of Science Part A, 48,* 52–59.
Haig, B. (2013). Analogical modeling: a strategy for developing theories in psychology. *Frontiers in Psychology, 4,* 1–3.
Hakli, R., Miller, K., & Tuomela, R. (2011). Two kinds of we-reasoning. *Economics and Philosophy, 26,* 291–320.
Harman, G. (1986). *Change in view. Principles of reasoning*. MIT Press.
Haslanger, S. (2000). Gender and race: (what) are they? (What) do we want them to be? *Noûs, 34*(1), 31–55.
Hausman, D. (2012). *Preference, value, choice, and welfare*. Cambridge University Press.
Heinonen, M. (2013). Tuomela's theory of the we-mode. In B. Kaldis (Ed.), *Sage encyclopedia of philosophy and the social sciences* (pp. 1053–1057). Sage.
Hempel, C. (1965). *Aspects of scientific explanation*. The Free Press.
Hesse, M. (1966). *Models and analogies in science*. University of Notre Dame Press.
Hofstadter, D. (2001). Analogy as the core of cognition. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *The analogical mind: perspectives from cognitive science*. MIT Press.
Holyoak, K. (2012). Analogy and relational reasoning. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). Oxford University Press.
Holyoak, K., & Thagard, P. (1997). The analogical mind. *American Psychologist, 52*(1), 35–44.
Humphreys, P. (2004). *Extending ourselves: computational science, empiricism, and scientific method*. Oxford University Press.
Hutto, D. (2008). *Folk psychological narratives. The sociocultural basis of understanding reasons*. MIT Press.
Jackson, F. (1998). *From metaphysics to ethics: a defense of conceptual analysis*. Oxford University Press.
Jacob, F. (1977). Evolution and tinkering. *Science, 196*(4295), 1161–1165.
Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Cambridge University Press.

Kermack, W., & McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A, 115*(772), 701–721.
King, J. (2016). Philosophical and conceptual analysis. In H. Cappelen, T. Gendler, & J. Hawthorne (Eds.), *The Oxford handbook of philosophical methodology* (pp. 249–261). Oxford University Press.
Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33,* 315–365.
Knobe, J., & Nichols, S. (2008). *Experimental philosophy*. Oxford University Press.
Ladyman, J., & Ross, D. (2007). *Every thing must go: metaphysics naturalized*. Oxford University Press.
Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in the History and Philosophy of Modern Physics, 41,* 253–262.
Levins, R. (1966). The strategy of model-building in population biology. *American Scientist, 54*(4), 421–431.
Levy, A., & Currie, G. (2015). Model organisms are not (theoretical) models. *British Journal for the Philosophy of Science, 66,* 327–348.
Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50,* 249–258.
Lewis, D. (1986). *Philosophical papers* (Vol. 2). Oxford University Press.
Lewontin, R. (1970). The units of selection. *Annual Review of Ecology and Systematics, 1,* 1–18.
List, C., & Pettit, P. (2011). *Group agents. The possibility, design and status of corporate agents*. Oxford University Press.
Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Machery, E. (2017). *Philosophy within its proper bounds*. Oxford Universiry Press.
Maibom, H. (2003). The mindreader and the scientist. *Mind and Language, 18*(3), 296–315.
Mäki, U. (2007). MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis, 70,* 29–43.
Markman, A., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 23,* 431–467.
Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model-building. *Synthese 170,* 169–190.
Mele, A. (2009). *Effective intentions*. Oxford University Press.
Menzies, P. (2010). Reasons and causes revisited. In M. de Caro, & D. Macarthur (Eds.), *Naturalism and normativity*, pp. 142–170. Columbia University Press.
Michael, J., & Szigeti, A. (2019). "The Group Knobe Effect": evidence that people intuitively attribute agency and responsibility to groups. *Philosophical explorations, 22*(1), 44–61.
Miller, G., Galanter, E., & Pribram, K. (1960). *Plans and the structure of behavior*. Henry Holt.
Neale, S. (1992). Paul Grice and the philosophy of language. *Linguistics and Philosophy, 15,* 509–559.
Nelson, N. (2018). *Model behavior: animal experiments, complexity, and the genetics of psychiatric disorders*. University of Chicago Press.
Nersessian, N. (1999). Model-based reasoning in conceptual change. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery*. Kluwer.
Nersessian, N. (2002). In the theoretician's laboratory. Thought experimenting as mental modeling. *Philosophy of Science, 2,* 291–301.
Nersessian, N., & Macleod, M. (2017). Models and simulations. In L. Magnani & T. Bertolotti (Eds.), *Springer handbook of model-based science* (pp. 119–136). Springer.
Niiniluoto, I. (1988). Analogy and similarity in scientific reasoning. In D. Helman (Ed.), *Analogical reasoning: perspectives of artificial intelligence, cognitive science, and philosophy* (pp. 271–299). Kluwer.
Niiniluoto, I. (1999). *Critical scientific realism*. Oxford University Press.
Noë. Alva. . (2004). *Action in perception*. MIT Press.
O'Brien, L. (2015). *Philosophy of action*. Palgrave Macmillan.
Odenbaugh, J., & Alexandrova, A. (2011). Buyer beware: Robustness analyses in economics and biology. *Biology & Philosophy, 26,* 757–771.
Orzack, S. H., & Sober, E. (1993). A critical assessment of Levins's the strategy of model building in population biology (1966). *The Quarterly Review of Biology, 68,* 533–546.
Parker, W. (2006). Understanding pluralism in climate modeling. *Foundations of Science, 11,* 349–368.
Paul, L. (2012). Metaphysics as modeling: the handmaiden's tale. *Philosophical Studies, 160,* 1–29.
Pettit, P. (1996). *The common mind. An essay on psychology, society and politics*. Oxford University Press.
Preyer, G., & Peter, G. (2017). *Social ontology and collective intentionality. Critical essays on the philosophy of Raimo Tuomela with his responses*. Springer.
Rakoczy, H. (2017). The development of individual and shared intentionality. In J. Kiverstein (Ed.), *The Routledge handbook of the philosophy of the social mind* (pp. 139–151). Routledge.

Reichenbach, H. (1938). *Experience and prediction. An analysis of the foundations and the structure of knowledge*. The University of Chicago Press.
Rodrik, D. (2015). *Why economics works, when it fails, and how to tell the difference*. Oxford University Press.
Ross, D. (2010). The economic agent: not human, but important. In U. Mäki (Ed.), *Elsevier handbook of philosophy of science, vol. 13: economics*. Elsevier.
Rovane, C. (1998). *Bounds of agency*. Princeton University Press.
Schelling, T. (1978). *Micromotives and macrobehavior*. W.V. Norton.
Schmid, H. (2014). Plural self-awareness. *Phenomenology and the Cognitive Sciences, 13,* 7–24.
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417–457.
Searle, J. (1983). *Intentionality*. Cambridge University Press.
Searle, J. (2010). *Making the social world: the structure of human civilization*. Oxford University Press.
Sellars, W. (1963). *Science, perception and reality*. Routledge.
Shiller, R. (2017). Narrative economics. *American Economic Review, 107*(4), 967–1004.
Simon, H. (1969). *The sciences of the artificial* (2nd ed.). MIT Press.
Smith, M. (1994). *The moral problem*. Blackwell.
Spellman, B., & Holyoak, K. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of personality and social psychology, 62,* 913–933.
Suppes, P. (1960). A comparison of the meaning and uses of models in mathematics and the formal sciences. *Synthese, 12*(2–3), 287–301.
Tarski, A. (1953). A general method in proofs of undecidability. In A. Tarski, A. Mostowski, & R. Robinson (Eds.), *Undecidable theories*. North Holland Publishing.
Thagard, P. (1992). *Conceptual revolutions*. Princeton University Press.
Thomson-Jones, M. (2005). Idealization and abstraction: a framework. In M. Thomson-Jones & N. Cartwright (Eds.), *Idealization XII: correcting the model* (pp. 173–217). Rodopi.
Tomasello, M. (2019). *Becoming human: a theory of ontogeny*. Harvard University Press.
Tuomela, R. (2013). *Social ontology: collective intentionality and group agents*. Oxford University Press.
Turner, S. (2018). *Cognitive science and the social*. Routledge.
Van Fraassen, B. (1980). *The scientific image*. Oxford University Press.
Varian, H. (2010). *Intermediate microeconomics: a modern approach* (8th ed.). W.V. Norton.
Velleman, D. (2009). *How we get along*. Cambridge University Press.
von Wright, G. H. (1971). *Explanation and understanding*. Routledge.
Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73,* 730–742.
Weisberg, M. (2007a). Who is a modeler? *British Journal for the Philosophy of Science, 58,* 207–233.
Weisberg, M. (2007b). Three kinds of idealization. *The Journal of Philosophy, 104*(12), 639–659.
Weisberg, M. (2013). *Simulation and similarity: using models to understand the world*. Oxford University Press.
Weisberg, M. (2016). Modeling. In H. Cappelen, T. Gendler, & J. Hawthorne (Eds.), *The Oxford handbook of philosophical methodology* (pp. 262–286). Oxford University Press.
Williamson, T. (2017). Model-building in philosophy. In R. Blackford & D. Broderick (Eds.), *Philosophy's future: the problem of philosophical progress* (pp. 159–172). Wiley Blackwell.
Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. Brewer & B. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–163). Jossey-Bass.
Winsberg, E. (2003). Simulated experiments: methodology for a virtual world. *Philosophy of Science, 70,* 105–125.
Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.
Wolfram, S. (2002). *A new kind of science*. Wolfram Media.
Woodward, J. (2003). *Making things happen: a theory of causal explanation*. Oxford University Press.
Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13,* 219–240.
Ylikoski, P. (2014). Agent-based simulation and sociological understanding. *Perspectives on science, 22*(3), 318–335.